iBinom Quick Start Guide

Contents

Introduction

Welcome to the iBinom genome data analysis service!

Aiming to get a foothold in the emerging clinical next-generation sequencing (NGS) market iBinom launched a simple, fast and affordable human genome interpretation service. iBinom SaaS platform is an end-to-end genome data analysis solution for diagnostics of Mendelian diseases. Unlike typical computer software, it is always up-to-date without the need to install updates. Support of non-Mendelian diseases, tumor profiling, and drug response biomarker discovery studies is coming soon.

iBinom uses Amazon EC2, the world's fastest proprietary algorithms, unique machine learning technology and special design of medical report to make inherited diseases diagnostics simpler and faster. In addition, iBinom has developed an interactive interface; clear and explicit for users with a modest bioinformatics expertise, it enables clinicians to easily operate the NGS data and detect the causative variants.

We believe that bringing iBinom service to medicine will solve the diagnostic and treatment odyssey for seriously ill patients.

This user manual will help you understand the iBinom service better. We are constantly updating the manual with the latest explanations on new features and updates to existing functionality. You can get the latest version by sending an email to info@ibinom.com or by simply downloading the manual from the iBinom website.

We wish you an informative read; should you have any further questions on using the iBinom service, please get back to us and we will be keen to assist you at any time.

Singing up for iBinom account

To start using the service, you should create your personal account. Please proceed with the following instructions:

Visit iBinom homepage at https://www.ibinom.com.

1. Click the blue button ”Try now”. You will see a window with a registration form.
2. Fill registration fields with your personal details. Communicate to us how you got aware about our service. If you are a Clinical Genomics Training Center’s trainee, tick a box next to the CGTC name.
3. Create your password; you will have to enter it each time you access the service. Once the registration form is completed, you will receive an account activation email.
4. Follow the link provided in the e-mail to complete your registration. You will be redirected to the iBinom website and logged in automatically.

Congratulations! Your personal iBinom account has been created. All the data submitted to your account including your personal details are secured and available only to the account holder. Please do not communicate your password to anyone else in order to keep your privacy. Once registered, you will be able to perform 6 panel or 6 exome VCF file analyses and 2 panel or 2 exome FASTQ/BAM file analyses free of charge.

Logging in to your account

You can always access your iBinom account from the homepage at https://www.ibinom.com by clicking the “Login” button. After entering your login and password, you will be directed to the Dashboard page of your account.

If you have forgotten your password, please click on "Forgot password" on the login window. A password reset instructions will be sent to your registered email address.

Should you have problems with accessing your account at any time, please send an email to info@ibinom.com and our Support team will get back to you as soon as possible.

Getting started

Once you have logged in to your iBinom account, you will see the Dashboard page shown as per Fig. 1.

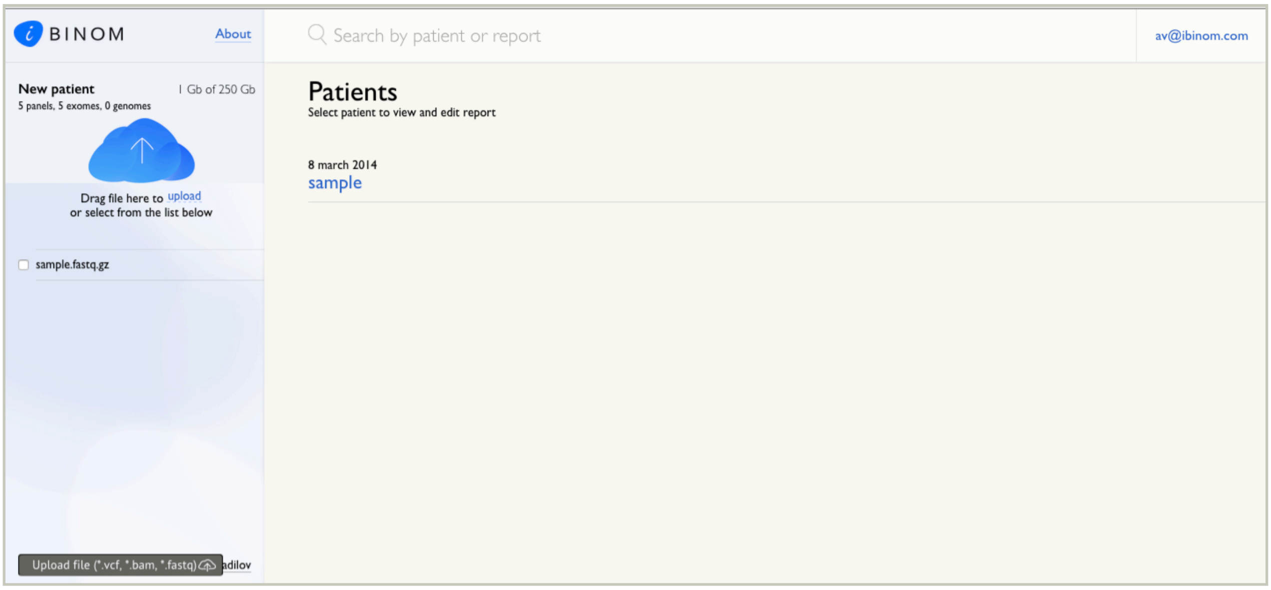

BINOM
About
Search by patient or report
av@ibinom.com
New patient
5 panels, 5 exomes, 0 genomes
1 Gb of 250 Gb
Drag file here to upload
or select from the list below
sample.fastq.gz
Upload file (*.vcf, *.bam, *.fastq)
Patients
Select patient to view and edit report
8 march 2014
sample

Fig. 1 iBinom Dashboard window

A preconfigured dataset called “sample” is available for you to get acquainted with the service. To have an overview of iBinom interface features, mouse to “sample” patient at the “Patients” list as shown on Fig. 2.

You will see “sample” highlighted with red color and various options you may select to proceed with. Click on a corresponding link at the right side to download the Quality Control report or variant call file in VCF/CSV format for demo data.

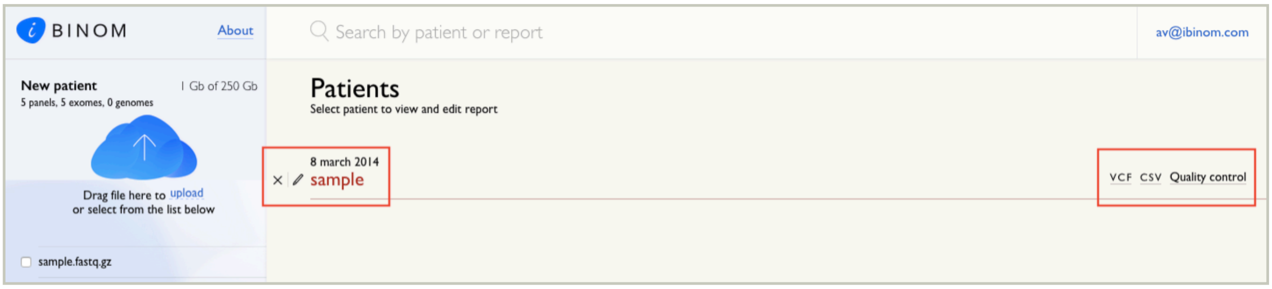
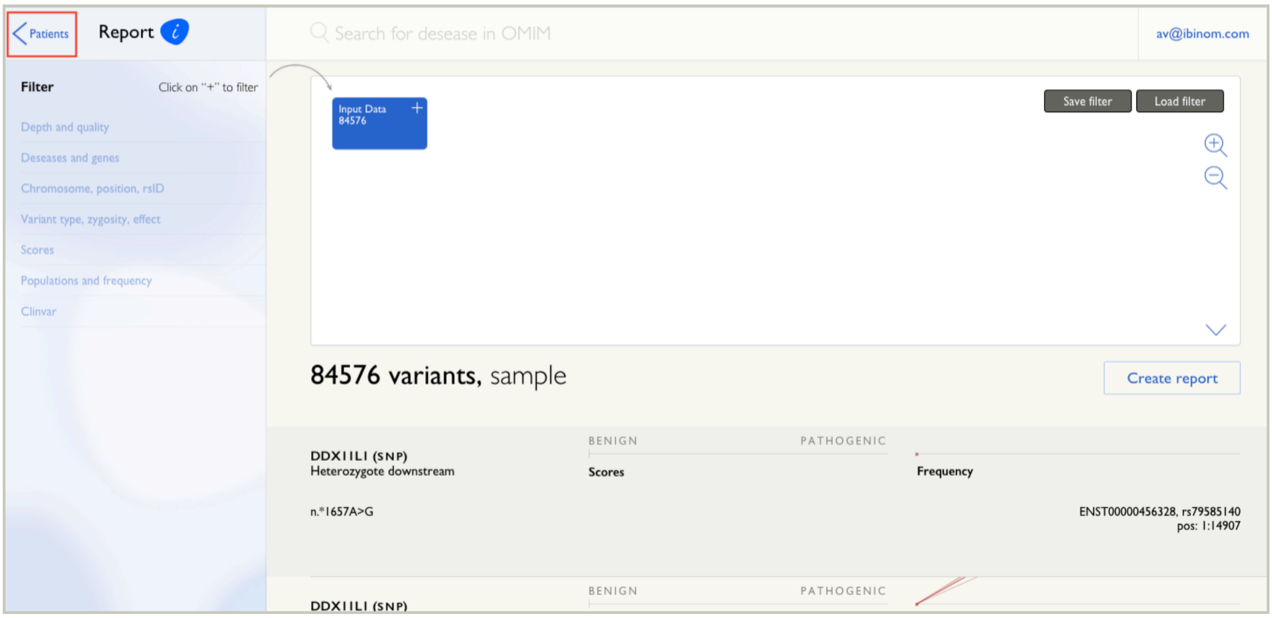


Fig. 2 Patient tab

Click on the Patient name to access the Filtering page and to view a list of variants.



84576 variants, sample

DDX11L1 (SNP)
Heterozygote downstream

n.*1657A>G

ENST00000456328, rs79585140
pos: 1:14907

DDX11L1 (SNP)

Fig. 3 Filtering page

To return to the Dashboard from the Filtering page, you can always click on the “Patients” button at the top left side of the page as you see on Fig. 3.

Using iBinom

Uploading your data

To upload your own sequencing data for analysis, click on the Cloud icon (see Fig. 4) on the left side of the Dashboard or drag and drop your files on it and confirm uploading.

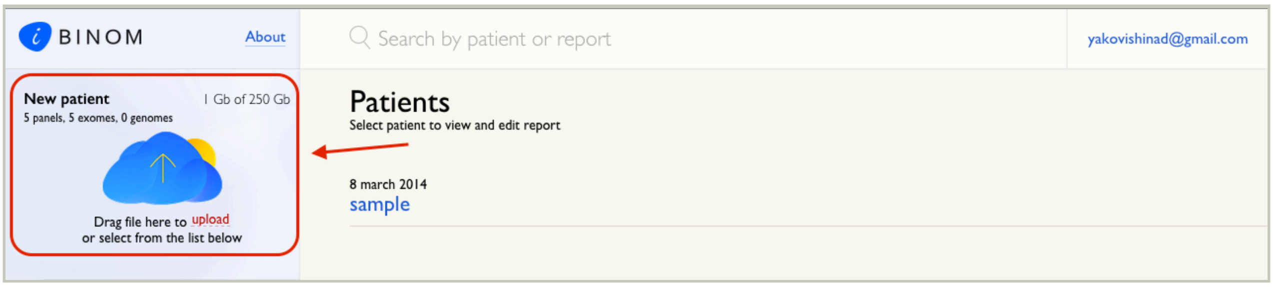

BINOM
About
Search by patient or report
yakovishinad@gmail.com
New patient
1 Gb of 250 Gb
5 panels, 5 exomes, 0 genomes
Drag file here to upload
or select from the list below
Patients
Select patient to view and edit report
8 march 2014
sample

Fig. 4 Uploading new files

Once uploading has started, an arrow indicator of loading process appears in the bottom of the screen.

At the moment iBinom supports raw sequencing data files in gzip compressed formats:

- *.fastq.gz
- *.fq.gz
- *.vcf.gz

and uncompressed formats:

- *.fastq
- *.fq
- *.bam
- *.vcf
- *.ab1

Note that while both compressed and uncompressed data are accepted, uploading of gzip formats takes less time, therefore, decreasing total amount of time required per an analysis.

Once uploading has completed, your files are displayed at the left bar of the Dashboard.

In order to delete the file, mouse to the corresponding file name and press the cross icon which will appear on the right side. ***Be careful as this operation cannot be undone.*** You may also download the file to your computer by clicking on the small cloud icon located next to the cross as per Fig. 5.

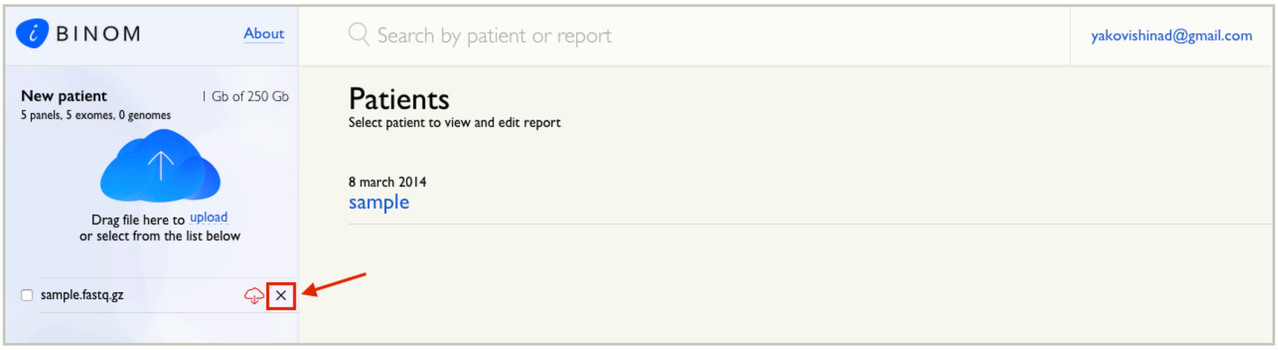

BINOM
About
Search by patient or report
yakovishinad@gmail.com
New patient
1 Gb of 250 Gb
5 panels, 5 exomes, 0 genomes
Drag file here to upload
or select from the list below
sample.fastq.gz
Patients
Select patient to view and edit report
8 march 2014
sample

Fig. 5 Removing a file

Starting data analysis

To start data analysis, follow the instructions and see Fig. 6.

1. Select a file from the left bar of the Dashboard.
2. Press "Create a new patient" button and confirm beginning of the analysis.

To see an analysis progress, check an indicator at the right side.

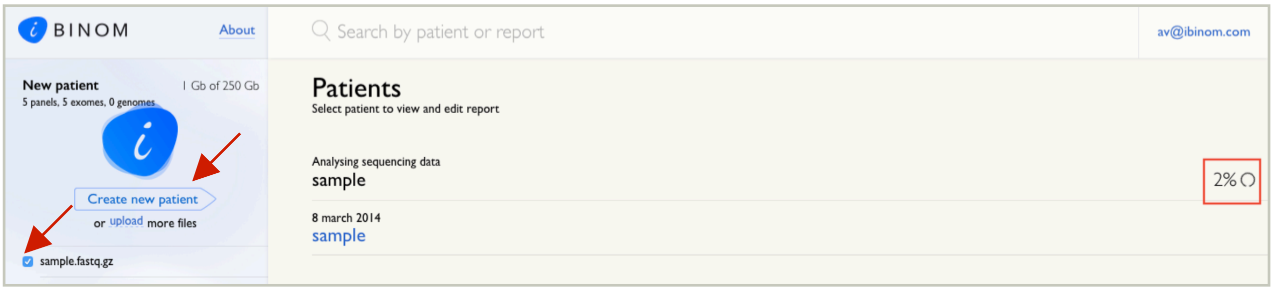
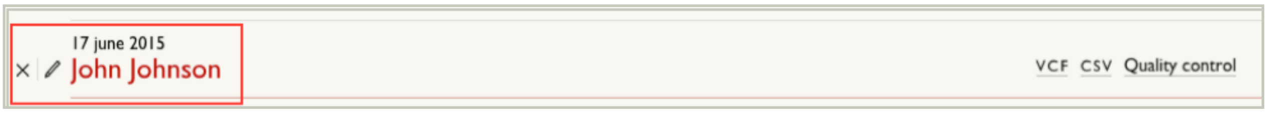


Fig. 6 Launching analysis

Accessing the results and filtering page

Once analysis has been completed, you may access the results and create a medical report. By default, a report will have the same name as the corresponding input file. You may delete the file or rename it by mousing over the patient and clicking cross or pencil button that will appear, respectively, as displayed on Fig. 7.



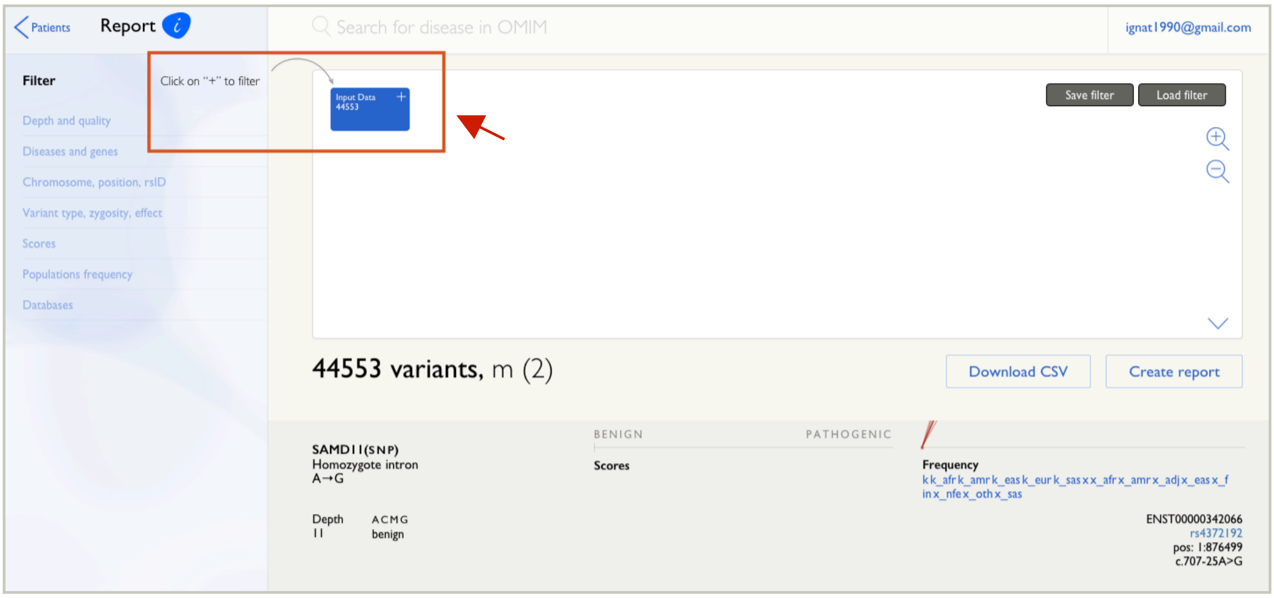
17 june 2015

John Johnson

Fig. 7 Renaming the file

To download VCF, CSV or Quality Control report corresponding to the sample, navigate and select the required option. To view and start variant filtering, simply click on the sample name – you will be redirected to the Filtering page. At the start, all the variants contained in VCF file are displayed on the Filtering page.

To filter the candidate variants from all the rest, you may take advantage of filtering features. On the Filtering page you will find a left bar with the filtering options along with the interactive filtering window. By clicking on "+" at the blue box as shown on Fig. 8 the left filtering bar is activated and you are able to select the options and sort variants of your interest.



44553 variants, m (2)

SAMD11(SNP)
Homozygote intron
A→G

Scores

Frequency
kk_afrk_amrk_eask_eurk_sasxx_afrx_amrx_adjx_easx_f inx_nfex_othx_sas

Depth 11

ACMG benign

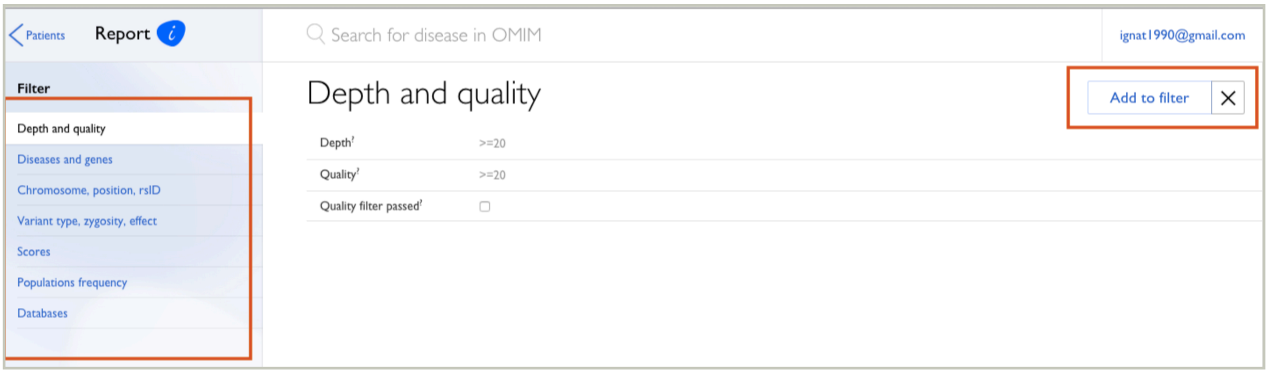
ENST00000342066
rs4372192
pos: 1:876499
c.707-25A>G

Fig. 8 Starting filtering

Filtering options

Once you have accessed the Filtering page, you may sort variants by the following options:

- Depth and quality
- Diseases and genes
- Chromosome, position, rsID
- Variant type, zygosity, effect
- Scores
- Population frequencies
- Databases



Depth' >=20
Quality' >=20
Quality filter passed' □

Fig. 9 Filtering bar

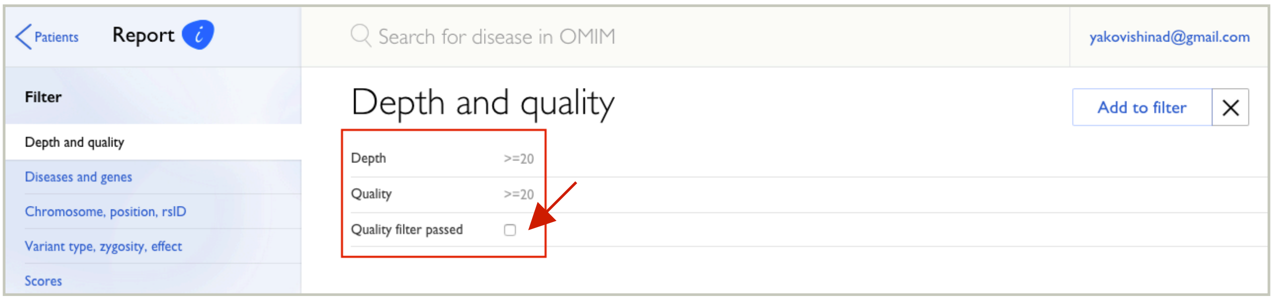
For each option you may select certain parameters. Please remember to click on the 'Add to filter' button to apply the selected parameters to the filter as shown on Fig. 9; otherwise the system will not proceed with it.

1. Selecting depth of coverage and mapping quality

Depth of coverage indicates how many times a given base was sequenced, i.e. how many reads cover the base. If a coverage value is low (in comparison with one allele coverage), this SNP might be generated because of sequencing or mapping errors and should be treated with caution.

Mapping quality indicates the minimum mapping quality of the variants that is accepted to the filtering process. The parameter shows how probable is the variant position. 20 means that read is mapped correctly with a probability of 99%.

To change depth or quality parameter mouse on “>=20” and type in a new value (see Fig. 10) If you unsure about the appropriate parameters tick the box next to “Quality filter passed”. Mouse on the field name, to see the exact preconfigured values the system applies to ensure the variant quality (ex. read depth >=10, p-value of biases >=0.05).



Depth and quality

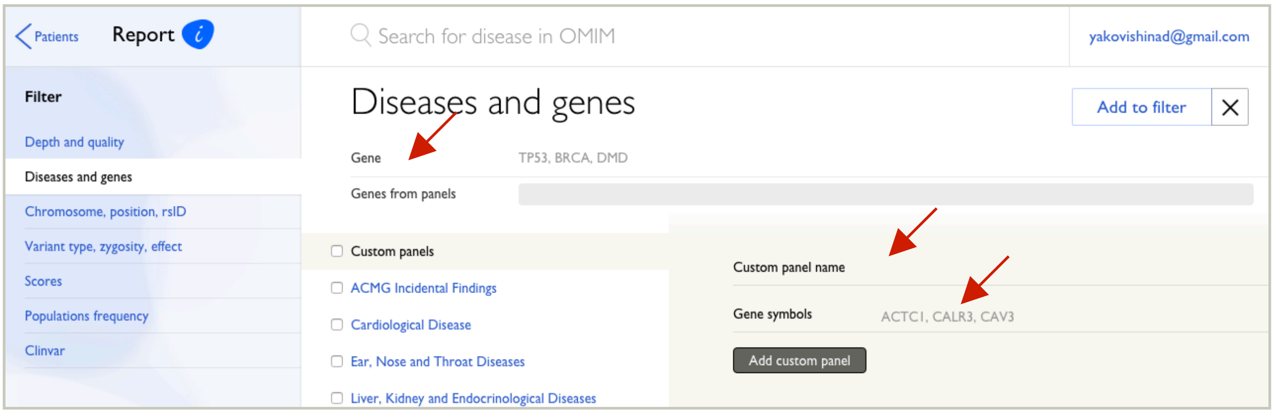
Depth	>=20
Quality	>=20
Quality filter passed	□

Fig. 10 Setting "Depth and quality" filtration parameters

2. Selecting diseases and gene panels (creating a custom panel)

Panel is the list of genes that have been associated with a given disease. Genomic panels were identified based on information in the Online Mendelian Inheritance in Man (OMIM) catalog2. To select the panels from the list, tick the boxes next the panels of interest.

To filter by genes of your interest, type a list of genes in the field “Gene”. You may create your own panel as explained on Fig. 11.



Diseases and genes

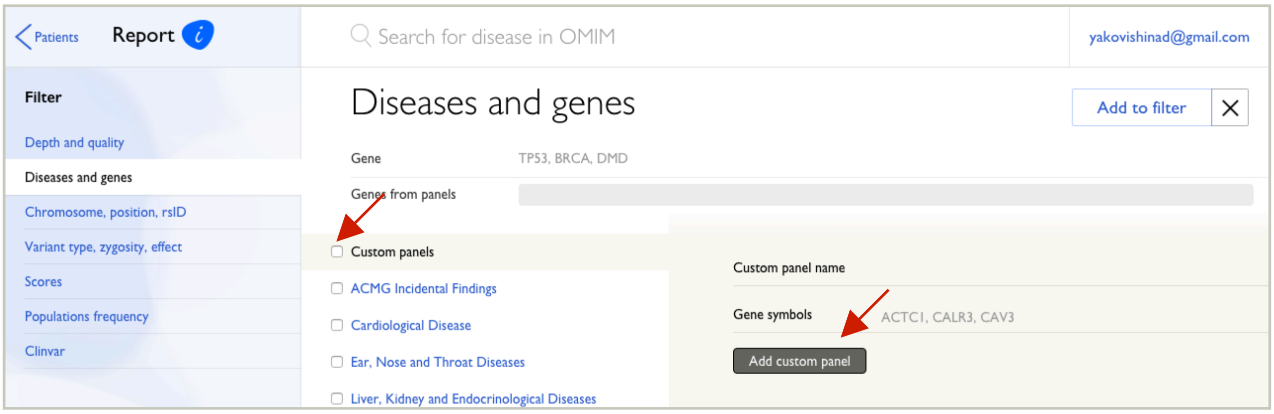
Gene	TP53, BRCA, DMD
Genes from panels	

- □ Custom panels
- □ ACMG Incidental Findings
- □ Cardiological Disease
- □ Ear, Nose and Throat Diseases
- □ Liver, Kidney and Endocrinological Diseases

Custom panel name	
Gene symbols	ACTC1, CALR3, CAV3

Fig. 11 Creating a custom panel

To add a custom panel , please name it and list genes separated with commas in the related fields, then, click the button “Add custom panel”. Your custom panel will appear on the screen as displayed on Fig. 12.



Patients · Report · Search for disease in OMIM · yakovishinad@gmail.com

Filter
- Depth and quality
- Diseases and genes
- Chromosome, position, rsID
- Variant type, zygosity, effect
- Scores
- Populations frequency
- Clinvar

Diseases and genes · Add to filter · ×

Gene: TP53, BRCA, DMD

Genes from panels

□ Custom panels
□ ACMG Incidental Findings
□ Cardiological Disease
□ Ear, Nose and Throat Diseases
□ Liver, Kidney and Endocrinological Diseases

Custom panel name

Gene symbols: ACTC1, CALR3, CAV3

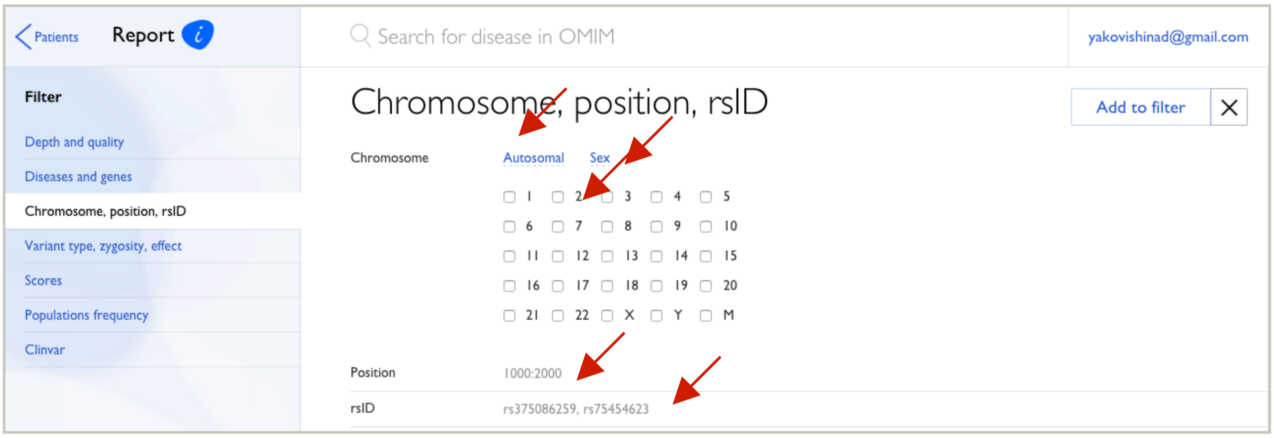
Add custom panel

Fig. 12 Selecting a custom panel

You may create as many custom panels as you wish. They all will be displayed on the screen once you go for "Custom panels". To delete a panel, click on a cross on the right side next to the panel name. In order to proceed with filtering, tick the box next to a created panel and press "Add to filter".

3. Selecting chromosomes, positions, rsID

The Tab 'Chromosomes, positions, rsID' enables choosing chromosomes, position or rsIDs (Rapid Stain Identification Series) of interest. Chromosomes can be chosen either separately or by groups: Autosomal (1-22) group and Sex (X and Y) group. In order to choose a certain chromosome position, at first, choose a chromosome and enter a position in the following format - 1000:2000. To specify an rsID, just enter list of comma-separated rsIDs to the corresponding field as shown on Fig. 13.



Patients · Report · Search for disease in OMIM · yakovishinad@gmail.com

Filter
- Depth and quality
- Diseases and genes
- Chromosome, position, rsID
- Variant type, zygosity, effect
- Scores
- Populations frequency
- Clinvar

Chromosome, position, rsID · Add to filter · ×

Chromosome: Autosomal · Sex

□ 1 □ 2 □ 3 □ 4 □ 5
□ 6 □ 7 □ 8 □ 9 □ 10
□ 11 □ 12 □ 13 □ 14 □ 15
□ 16 □ 17 □ 18 □ 19 □ 20
□ 21 □ 22 □ X □ Y □ M

Position: 1000:2000

rsID: rs375086259, rs75454623

Fig. 13 Setting filtration parameters for chromosome, position and rsID

4. Selecting variant type, zygosity and effect

Variant has three main *types*: SNP (single nucleotide polymorphism), DEL (deletion up to 10-15 bases) and INS (insertion up to 10-15 bases). You may choose one or several types at once (all types by default).

Zygosity – a property of the variant that shows if it is homozygous (variant is presented on two alleles) or heterozygous (variant is presented on one allele).

Effect – shows how the variant can effect on the protein behavior.

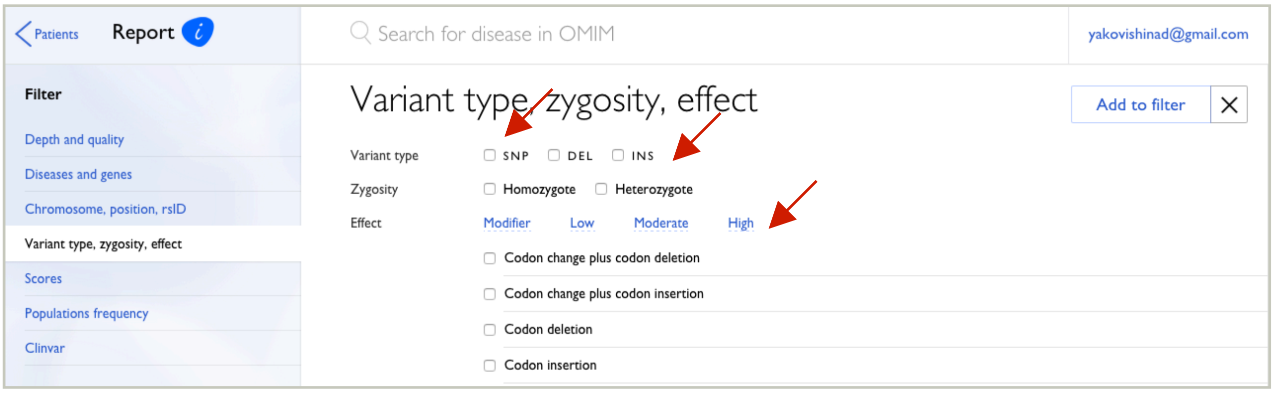
HIGH - the variant is assumed to have high (disruptive) impact on the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay. stop_gained, frameshift_variant.

MODERATE - a non-disruptive variant that might change protein effectiveness. missense_variant, inframe_deletion

LOW - assumed to be mostly harmless or unlikely to change protein behavior. synonymous_variant.

MODIFIER - usually non-coding variants or variants affecting non-coding genes, when predictions are difficult to make or there is no evidence of impact. exon_variant, downstream_gene_variant.

Select desired parameters by ticking the boxes next to them as per Fig. 14.



Variant type, zygosity, effect

Variant type □ SNP □ DEL □ INS

Zygosity □ Homozygote □ Heterozygote

Effect Modifier Low Moderate High

□ Codon change plus codon deletion

□ Codon change plus codon insertion

□ Codon deletion

□ Codon insertion

Fig. 14 Setting filtration parameters for variant type, zygosity and effect

5. Selecting scores

Scores - shows a probability that a variant affects the protein function; scores are given to variants by different tools. The tools included in the system are:

SIFT (Sorting Intolerant From Tolerant) predicts whether an amino acid substitution affects protein function. A score <=0.05 is considered damaging.

PolyPhen-2 (Polymorphism Phenotyping v2) predicts the possible impact of amino acid substitutions on the stability and function of human proteins. Probability for a variant being damaging: [0.909:1] - damaging; [0.447:0.908] - possibly damaging;

[0:0.446] - benign.

MutationTaster employs a Bayes classifier to eventually predict the disease potential of an alteration.The probability value is the probability of the prediction, i.e. a value close to 1 indicates a high 'security' of the prediction.

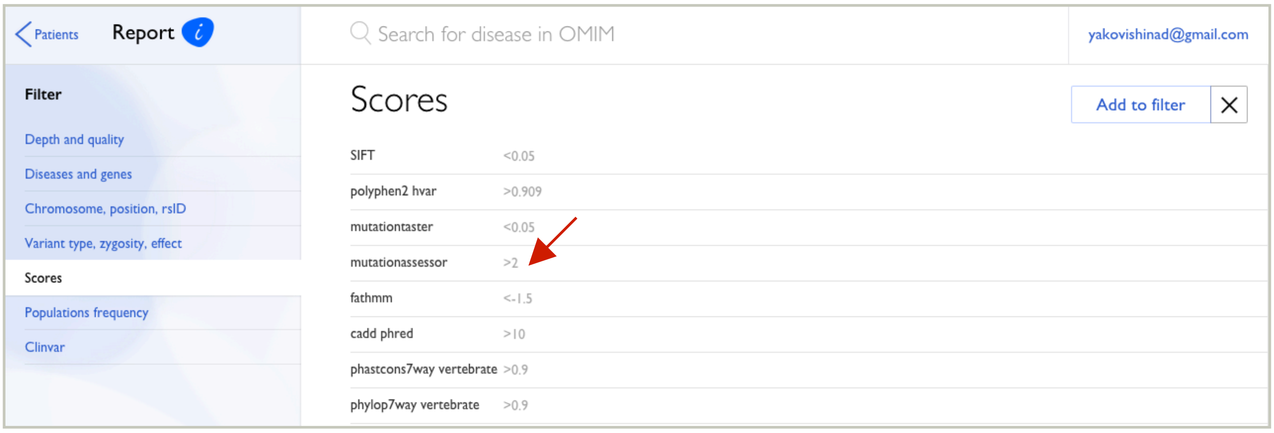
Mutationassessor predicts the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms. The lower the score, the more likely the SNP has damaging effect.

Fathmm - tool for predicting the functional effects of protein missense mutations. Less score is more pathogenic. Prediction "D(AMAGING)" for scores less than -1.5; otherwise -- "T(OLERATED)". For isoforms the smallest score (most damaging) was used. Range [-16.13 : 10.64].

Phastcons7way vertebrate - phylogenetic conservation score based on the multiple alignments: - log_10(p-value). The larger the score, the more conserved the site.

Phylop7way vertebrate - Phylogenetic conservation score based on the multiple alignments: - log_10(p-value). The larger the score, the more conserved the site.

To apply scoring parameters, type in a new value next to the score of interest as on Fig. 15 and click on ‘Add to filter’ button.



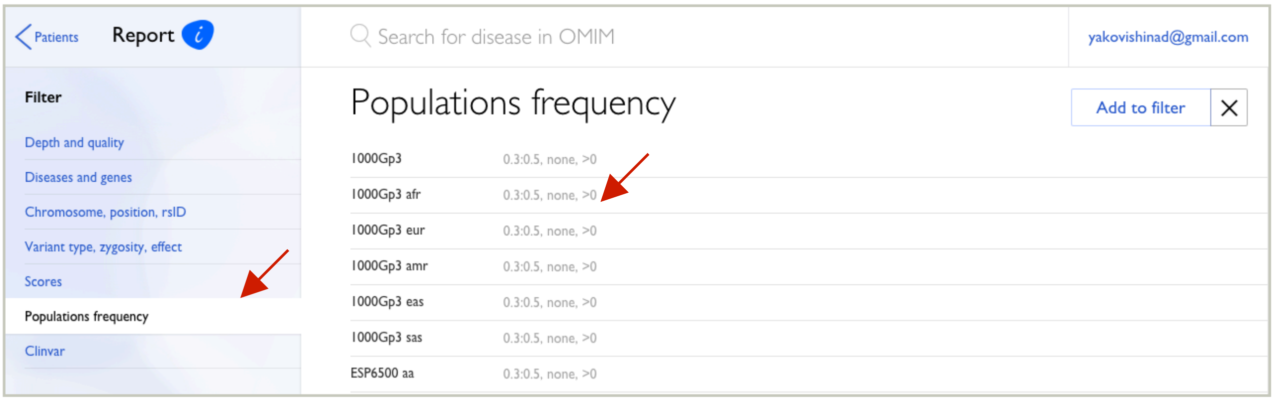
Scores

SIFT	<0.05
polyphen2 hvar	>0.909
mutationtaster	<0.05
mutationassessor	>2
fathmm	<-1.5
cadd phred	>10
phastcons7way vertebrate	>0.9
phylop7way vertebrate	>0.9

Fig. 15 Setting filtration parameters for scores

6. Selecting population frequency

Population frequency means how frequent a variant is represented in a population, i.e. a portion of people from a given population sharing the same variant. The less frequent a variant is in the population, the more chances that the variant is causative. To set the parameters, simply type the values in the fields as indicated on Fig. 16.



Populations frequency

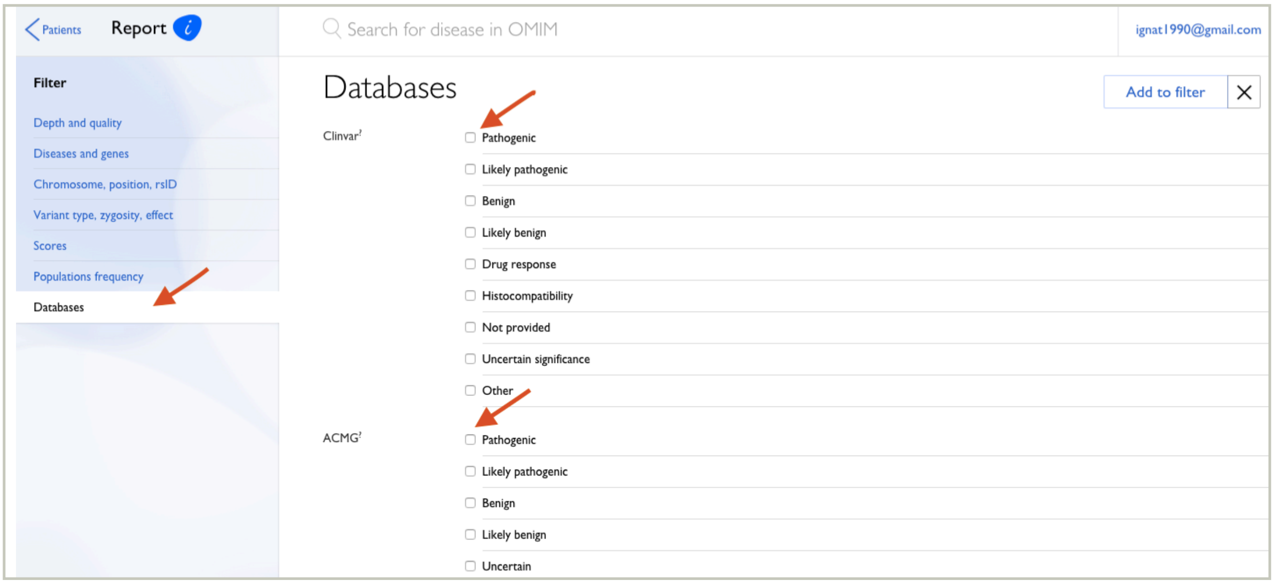
1000Gp3	0.3:0.5, none, >0
1000Gp3 afr	0.3:0.5, none, >0
1000Gp3 eur	0.3:0.5, none, >0
1000Gp3 amr	0.3:0.5, none, >0
1000Gp3 eas	0.3:0.5, none, >0
1000Gp3 sas	0.3:0.5, none, >0
ESP6500 aa	0.3:0.5, none, >0

Fig. 16 Setting filtration parameters for population frequency

7. Selecting database parameters

Clinvar is a database that aggregates supporting evidence of relationship among medically important variants and phenotypes. Variant classification varies from benign to pathogenic; a variant may be classified as of unknown significance when the supporting evidence is weak. Tick the box next to a desirable classification parameter as per Fig. 17 and click ‘Add to filter’.

Choose a parameter corresponding to American College of Medical Genetics’ recommendations (see http://www.ncbi.nlm.nih.gov/pubmed/25741868).



Databases

Clinvar[?]
- ☐ Pathogenic
- ☐ Likely pathogenic
- ☐ Benign
- ☐ Likely benign
- ☐ Drug response
- ☐ Histocompatibility
- ☐ Not provided
- ☐ Uncertain significance
- ☐ Other

ACMG[?]
- ☐ Pathogenic
- ☐ Likely pathogenic
- ☐ Benign
- ☐ Likely benign
- ☐ Uncertain

Fig. 17 Setting filtration parameters for Clinvar database

General filter use

Once you have chosen one or several parameters of filtering options, press button 'Add to filter'. You will be redirected to the interactive filter window (see Fig. 18 showing the filter adjusted for *Clinvar* option):

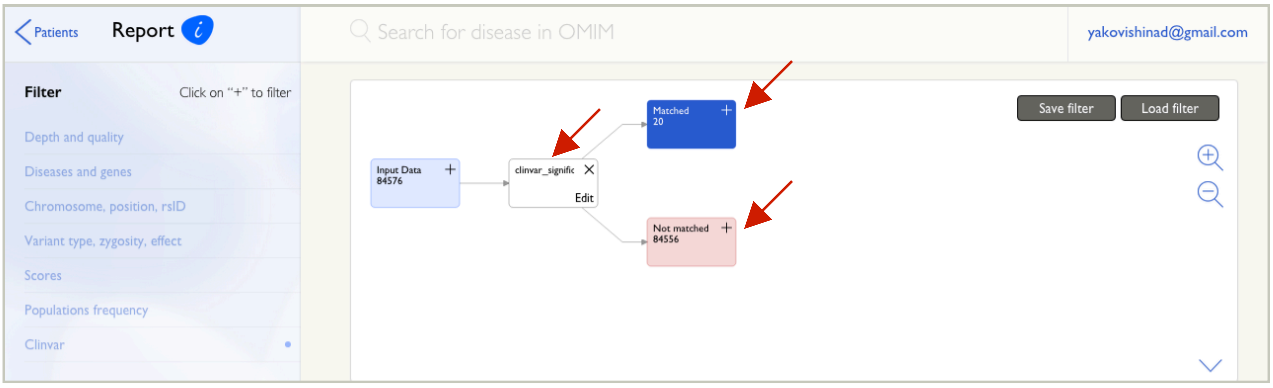


Fig. 18 iBinom interactive filter

In the white box a selected filtering option is displayed. In the blue box, the number of variants matching to your filter parameters is shown. The red box displays the number of variants that are not matched. You may add the next filtering option at any step: for matched, unmatched and initial input data, by clicking '+' inside the corresponding box. To remove a branch, click on 'x' inside the white box. At the end, you may build 'a tree' of the filtering options with you own parameters as shown on Fig. 19.

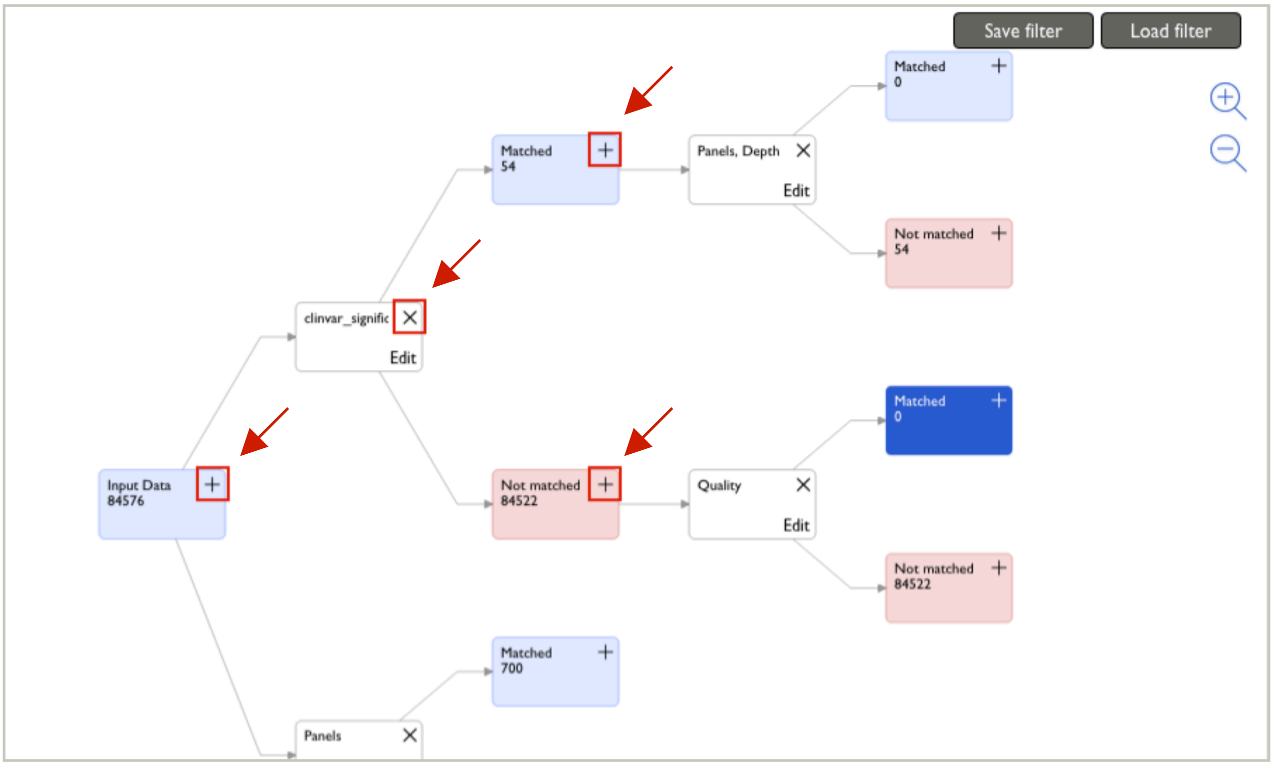
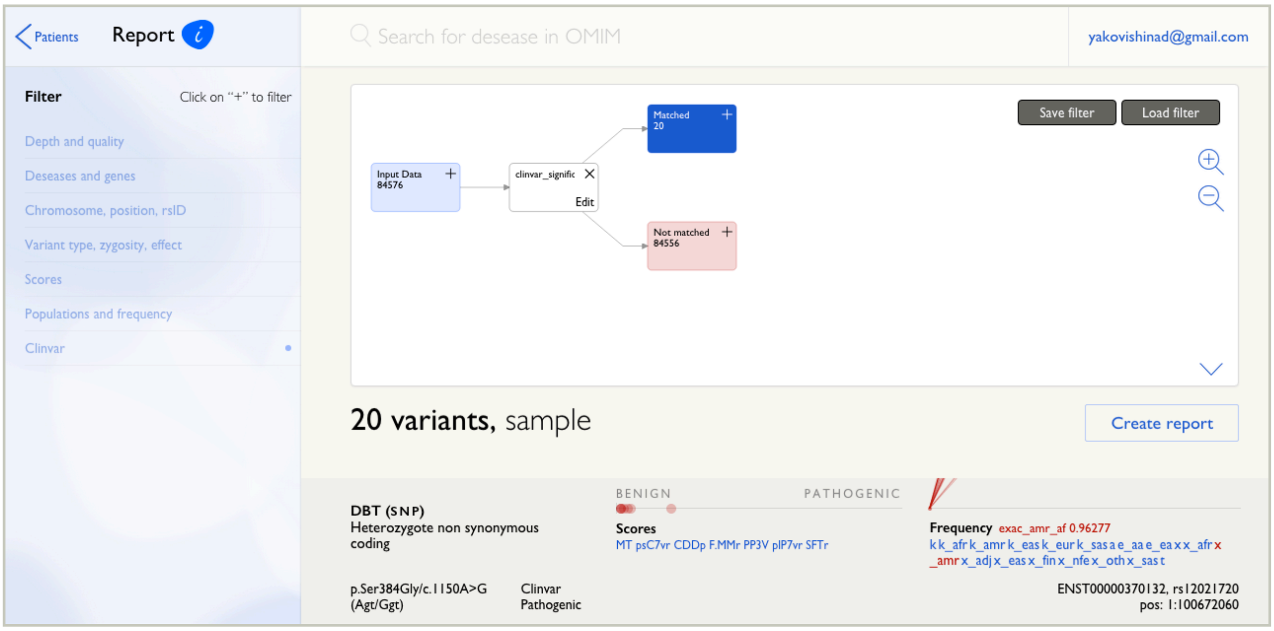


Fig. 19 iBinom extended interactive filter

Viewing filter results

To see the variants that passed a certain filtration step, click on a box of interest - the variants will be listed on the page below the interactive filter window as per Fig. 20.



20 variants, sample

DBT (SNP)
Heterozygote non synonymous coding

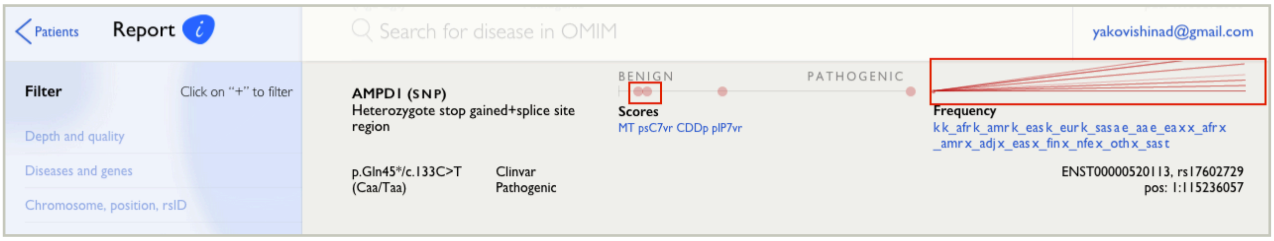
BENIGN PATHOGENIC
Scores
MT psC7vr CDDp F.MMr PP3V plP7vr SFTr

Frequency exac_amr_af 0.96277
kk_afr k_amr k_eas k_eur k_sas ae_aa e_ea x x_afr x_amr x_adj x_eas x_fin x_nfe x_oth x_sas t

p.Ser384Gly/c.1150A>G (Agt/Ggt)
Clinvar Pathogenic
ENST00000370132, rs12021720
pos: 1:100672060

Fig. 20 Viewing the results

The results comprise a list of variants that satisfy your filtering parameters. Fig. 21 displays the key elements of an example variant. Each variant contains information about its type, gene, location, score ranking from different tools and frequency in populations. Each score is represented by a circle on the 'Scores' scale, the left side of the scale corresponds to benign and the right - to pathogenic scores. To view the exact values of each score, mouse on the corresponding circles. List of frequencies is represented by a right graph. Each frequency is represented by a line, the greater the angle, the greater is the population frequency. Mouse on a line or on a population symbol, to see the exact frequency value.



AMPD1 (SNP)
Heterozygote stop gained+splice site region

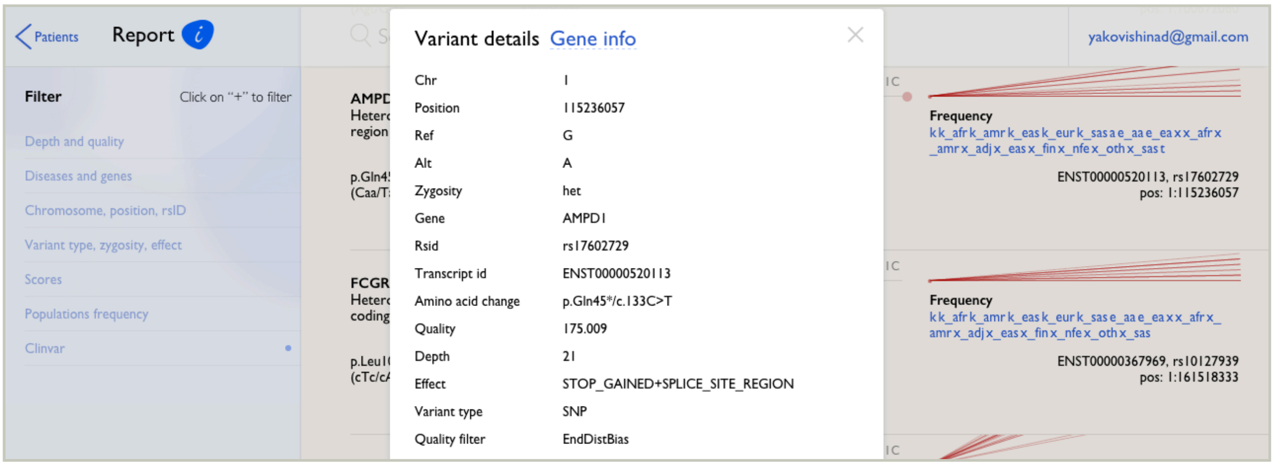
BENIGN PATHOGENIC
Scores
MT psC7vr CDDp plP7vr

Frequency
kk_afr k_amr k_eas k_eur k_sas ae_aa e_ea x x_afr x_amr x_adj x_eas x_fin x_nfe x_oth x_sas t

p.Gln45*/c.133C>T (Caa/Taa)
Clinvar Pathogenic
ENST00000520113, rs17602729
pos: 1:115236057

Fig. 21 Results page key elements

To see the variant and gene details, click on the variant. All available information will be shown in the pop-up window as on Fig. 22.


Patients
Report
Filter
Click on "+" to filter
Depth and quality
Diseases and genes
Chromosome, position, rsID
Variant type, zygosity, effect
Scores
Populations frequency
Clinvar
yakovishinad@gmail.com
Frequency
kk_afrk_amrk_eask_eurk_sasae_aae_eaxx_afrx_amrx_adjx_easx_finx_nfex_othx_sast
ENST00000520113, rs17602729
pos: 1:115236057
Frequency
kk_afrk_amrk_eask_eurk_sase_aae_eaxx_afrx_amrx_adjx_easx_finx_nfex_othx_sas
ENST00000367969, rs10127939
pos: 1:161518333

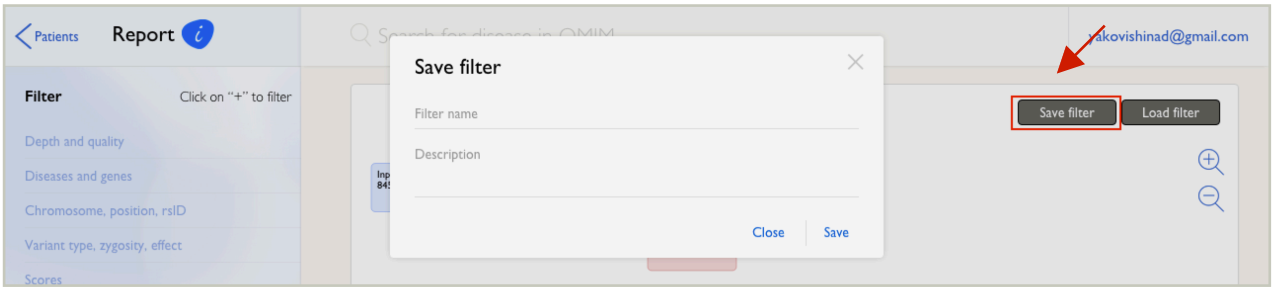
Variant details Gene info

Chr	1
Position	115236057
Ref	G
Alt	A
Zygosity	het
Gene	AMPD1
Rsid	rs17602729
Transcript id	ENST00000520113
Amino acid change	p.Gln45*/c.133C>T
Quality	175.009
Depth	21
Effect	STOP_GAINED+SPLICE_SITE_REGION
Variant type	SNP
Quality filter	EndDistBias

Fig. 22 Viewing variant details

Filter saving and variant report creating

To remember your filtration parameters, click on the grey button 'Save filter', fill the filter name and description (optional) fields in the pop-up window and press 'Save' as shown on Fig. 23. Thus, you will be able to apply at one click the same filtration parameters for your consecutive analyses.



Save filter

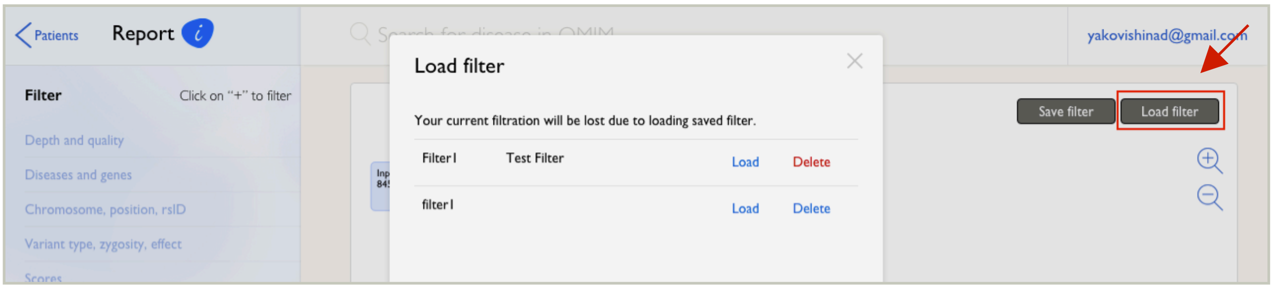
Filter name

Description

Close Save

Fig. 23 Saving the filter

To *load* the filter, simply click 'Load filter' and choose your filter as on Fig. 24.



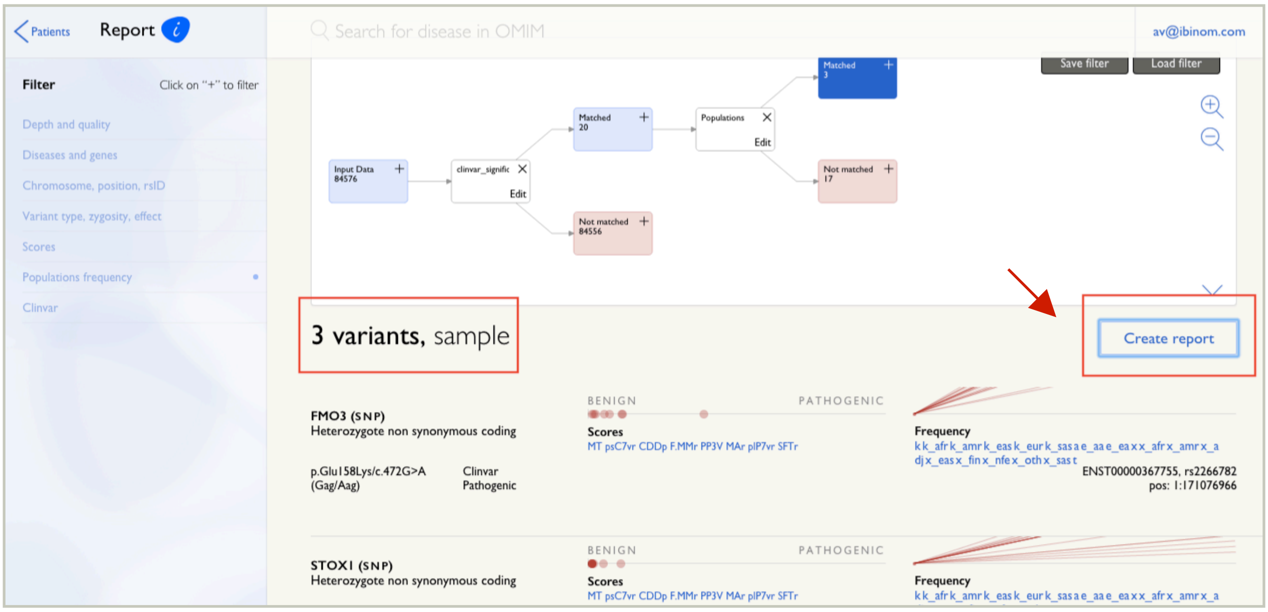
Load filter

Your current filtration will be lost due to loading saved filter.

Filter1	Test Filter	Load	Delete
filter1		Load	Delete

Fig. 24 Loading the filter

In order to *download* the most relevant variants remained after applied filtration steps, the system enables you to generate the iBinom variant report. Click on ‘Create report’ button as shown on Fig. 25 and the system will automatically save and open the PDF report file.



3 variants, sample

FMO3 (SNP)
Heterozygote non synonymous coding

p.Glu158Lys/c.472G>A (Gag/Aag)
Clinvar Pathogenic

BENIGN PATHOGENIC
Scores
MT psC7vr CDDp F.MMr PP3V MAr pIP7vr SFTr

Frequency
k k_afr k_amr k_eas k_eur k_sas a e_aa e_ea x x_afr x_amr x_a dj x_eas x_fin x_nfe x_oth x_sas t

ENST00000367755, rs2266782
pos: 1:171076966

STOX1 (SNP)
Heterozygote non synonymous coding

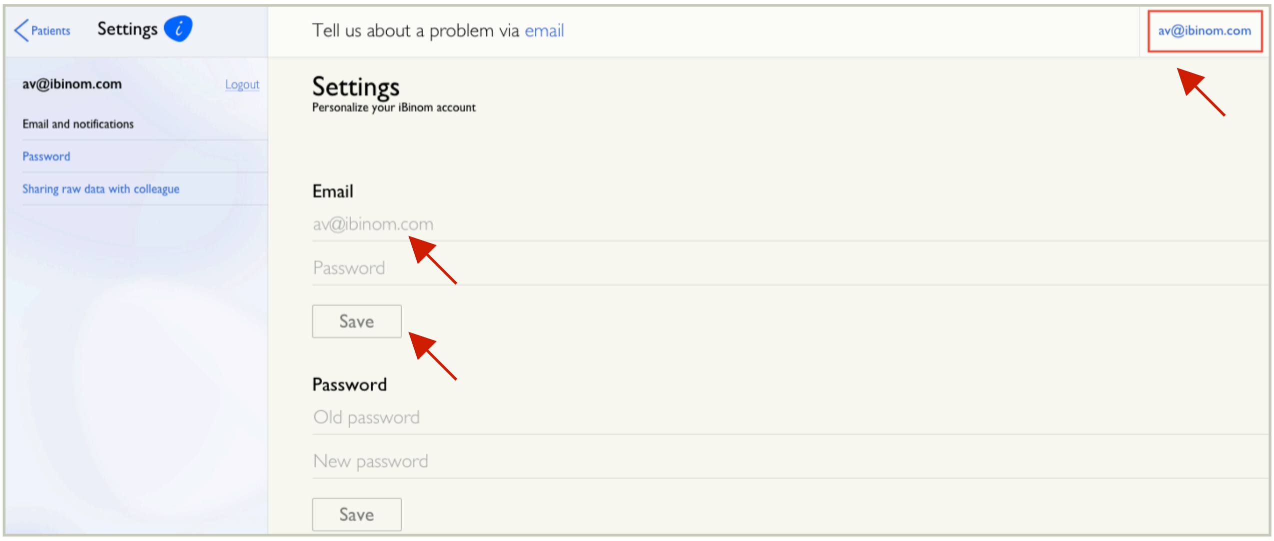
BENIGN PATHOGENIC
Scores
MT psC7vr CDDp F.MMr PP3V MAr pIP7vr SFTr

Frequency
k k_afr k_amr k_eas k_eur k_sas a e_aa e_ea x x_afr x_amr x_a

Fig. 25 Creating a variant report

Personal account settings

In order to access your personal account settings, click to your account e-mail shown in the top-right corner of the screen. Type in a new e-mail or password if you wish to modify them and click on the button ‘Save’ to apply the changes (Fig. 26).



Tell us about a problem via email

Settings
Personalize your iBinom account

Email
av@ibinom.com
Password

Password
Old password
New password

Fig. 26 Modifying the account settings

iBinom analysis technical details

General design

Sequencing data analysis consists of 5 steps:

1. Preprocessing
2. Reads alignment
3. Variants calling
4. Variants annotation
5. Report creating

Data uploaded to the server are protected with HTTP Secure (https) protocol.

We use the human genome reference version GRch37.75 (hg19) on each step of data analysis.

Preprocessing

Before the alignment process, we calculate fragments quality distribution.

Low-quality reads are filtered out from later processing.

The following statistics are calculated on filtered reads:

- Distribution of nucleotides phred-qualities
- Average quality per position within read
- Distribution of reads lengths
- Amount of each nucleotide in data

All statistics are provided in the iBinom Quality Report.

Fragments alignment

To align the reads, iBinom uses BWA-MEM tool adjusted to run on MapReduce cluster of Elastic MapReduce Amazon system. The BWA-MEM alignment algorithm is based on Burrows-Wheeler transformation. Specifically, seeds are extracted from the reads and aligned to the genome with the maximal exact matches. These seed alignments are extended with the affine-gap Smith-Waterman algorithm.

Alignment benchmarking

We have benchmarked the most applied open source read mappers. The study employed the public server for genome comparison and analytic testing GCAT http://www.bioplanet.com/gcat. Benchmarking has been performed on four single-end testing datasets offered by GCAT server. BWA-MEM tool proved to be the most efficient and accurate mapper. Results are represented in the table below.

Name	Total Correct Reads (%)	Incorrectly Mapped Reads (%)	Unmapped Reads (%)
BWA-MEM	7863529	101527	7
	98.73%	1.27%	0.00%
Bowtie2	7670364	251234	41901
	96.32%	3.15%	0.53%
BWA	7363467	81561	518471
	92.47%	1.02%	6.51%

Variant calling

Before starting variant calling, alignment filtering is applied. First, non-mapped reads and reads with low mapping quality are filtered out. Then, all remaining reads are grouped. Two reads are connected in one group if they intersect and this connection is transitive between the reads. When all reads are grouped, the obtained groups are sorted out by the maximum coverage depth in a group, by amount of reads and amount of covered nucleotides within a group. The step helps to remove low-covered regions preventing from false positive variants. Also this step filters out clusters that are not located within the regions of interest for targeted or exome sequencing.

Variant calling is performed using SAMtools libraries with tuned settings and adjusted to be run on MapReduce cluster. SAMtools algorithm uses Bayesian model to detect variants and its quality values. Local realignment is performed to recover indels that occur at the end of a read but appear to be contiguous mismatches.

Variant calling benchmarking

We have benchmarked SAMtools (see as iBinom on diagram) against the most popular open source variant calling pipelines Bowtie2 + GATK and BWA + samtools, and proprietary DNANexus solution using GCAT service. We used Illumina Exome 30x covered NA12878 dataset as we considered it the typical data that one could upload to our solution. The dataset is widely perceived to be the "gold standard" as a part of wet lab and bioinformatics pipeline validation.

The Venn diagram results are represented on Fig. 27.

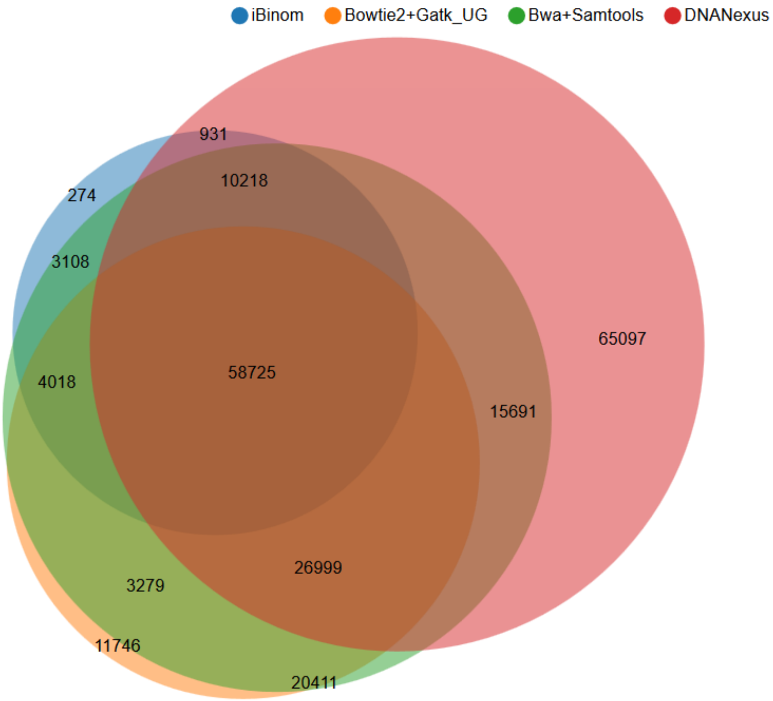


Fig. 27 Concordance of variants called by multiple pipelines

Variants called by iBinom solution demonstrated the top level of concordance compared to the other pipelines due to applied filtering algorithms. High discordance of the variants detected by the other pipelines occurs in some positions and might be explained by varying features and thresholds that variant callers use.

Because of the low coverage, such positions are supposed to produce more errors in variant calling too. As soon as for medical applications we seek for the most confident variants, iBinom prioritizes a higher precision rate over sensitivity and specificity.

To measure the calling quality, we used genotyping data from HumanOmni2.5-8v1 array provided by GCAT server.

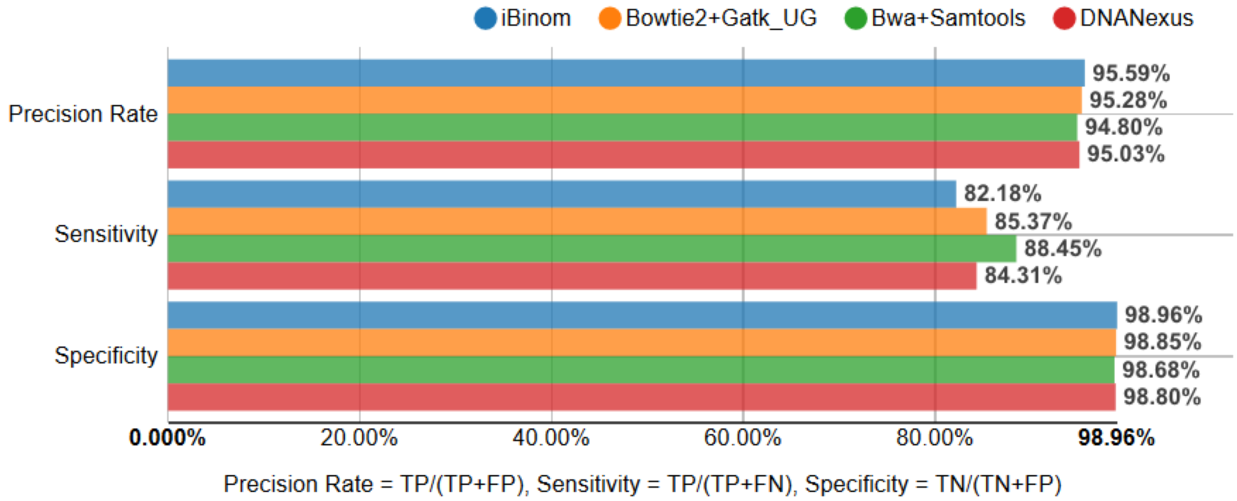


Fig. 28 Quality of predicted variants

As one may see on Fig. 28, the iBinom solution demonstrates results comparable to the widely applied solutions and even outperforms them in Precision rate and Specificity.

Annotation

Annotation is performed in two key steps.

First, we annotate the variants followed by prediction of their coding effects on known genes such as synonymous or non-synonymous SNPs, start codon gains or losses, stop codon gains or losses.

Also for each variant we evaluate its genomic locations within intronic, 5' UTR, 3' UTR, upstream, downstream or intergenic regions.

Second, the iBinom machine learning algorithm is used to define the level of variants pathogenicity called “iBinom score”. The iBinom score varies from 0 to absolutely non-pathogenic mutation to 1 for extremely pathogenic mutation. All variants with the iBinom score exceeding 0,9 should be considered as potentially clinically significant.

For the purpose of algorithm training procedure, we used information from the databases and predictors listed below. The most generally recognized and reputed scores values and databases information are represented on the Filtering Results page and included in the iBinom variant report. iBinom continue improving the service by updating it with additional databases and features. ***Stay tuned!***

List of the external annotation databases and scores the iBinom service refers to:

- 1000 Genomes, phase 3, 15’ feb
- Clinvar, 15’ sep
- dbSNP build 144, 15’ jun
- dbNSFP, v3.0, 15’ aug
- snpEff 4.1b, 15’ feb
- MutationAssessor, v2, 13’ sep
- MutationTaster, v2, 14’ jul
- ExAC v0.3, dec’ 14
- ESP6500, 14’ nov
- UK10K, 15’ jun
- PolyPhen hvar, v 2.2.2, 12’ feb
- Fathmm, v2.3, 14’ nov
- SIFT, ensembl 66, 15’ jan
- phastCons, 14’ jun
- phyloP 14’ jun
- OMIM, 15’ oct
- SiPhy 29way, 11’ may
- GERP++ , 11’ may